基于子话题分治匹配的新事件检测

洪 宇 张 宇 范基礼 刘 挺 李 生
(哈尔滨工业大学计算机科学与技术学院信息检索研究室 哈尔滨 150001)

摘 要 新事件检测是话题检测与跟踪领域的一项重要研究,其任务是实时监控新闻报道流并从中识别新话题.现有方法将话题和报道描述为单一结构的特征向量进行匹配,造成子话题间互为噪声并形成错误语义,从而误导新话题的识别.针对这一缺陷,文中提出基于子话题分治匹配的新事件检测方法,将话题和报道划分为不同子话题,根据相关子话题的比例关系和分布关系建立新话题识别模型.实验在 TDT4 和 TDT5 中获得显著改进,最小检测错误代价为 0.4061,相应漏检率为 0.1859.

关键词 新事件检测;话题检测与跟踪;子话题
中图法分类号 TP391

New Event Detection Based on Division Comparison of Subtopic

HONG Yu ZHANG Yu FAN Ji-Li LIU Ting LI Sheng
(*Information Retrieval Laboratory, School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract New event detection is an important research in the field of topic detection and tracking, and its task is real-time monitoring the stream of news stories and identifying the new topics in it. Current methods match the topics and stories as they are single-structured vectors of terms, which make the subtopics become noises of each other, and these noises often describe wrong semantics, by which the identification of new topics would be misled. In response to this defect, this paper proposes a new event detection method based on division comparison of subtopics, which divided each topic and story into different subtopics and identified new topic basing on the proportion and distribution relations of the relevant subtopics. This method achieves substantial improvement on TDT4 and TDT5, whose minimum cost of detection error is 0.4061 and missing probability is 0.1859.

Keywords new event detection; topic detection and tracking; subtopic

1 引 言

话题检测与跟踪(Topic Detection and Tracking, TDT)是一项针对新闻报道进行信息识别、挖掘和组织的研究,其为信息检索、数据挖掘和自然语言处理等技术提供了全新的多语言测试平台[1].新事件检测(New Event Detection, NED)是 TDT 中的一项重要子任务,其定义为检测时序新闻流中对某一话题的首次报道[2],即识别新话题.话题[3]由一个种子事件以及后续直接相关的事件或活动组成,子话题是针对其中某一事件的相关描述,事件则定义为发生于特定时间和特定地点的事情[4].例如,"2001 年 9 月 11 日针对美国世贸和五角大楼的恐怖袭击"是话题"911"的种子事件,它与"灾后处理"、"嫌疑犯调查"和"国际社会援助"等后续相关事件构

收稿日期:2007-12-10.本课题得到国家自然科学基金(60435020, 60503072)与国家"八六三"高技术研究发展计划探索类专题项目(2006AA01Z145)资助.洪 宇,男,1978 年生,博士研究生,主要研究方向为话题检测与跟踪、信息过滤、个性化信息检索.E-mail: hy@ir.hit.edu.cn.张 宇,男,1972 年生,博士,副教授,主要研究方向为信息过滤、自动问答、自然语言处理.范基礼,男,1986 年生,硕士研究生,研究方向为话题检测与跟踪.刘 挺,男,1972 年生,博士,教授,博士生导师,研究方向为自然语言处理、信息检索.李 生,男,1943 年生,博士,教授,博士生导师,主要研究领域为自然语言处理,信息检索与机器翻译.

成完整的“911”话题,其中对每个真实事件的相关描述构成了该话题内的不同子话题.

通过 NED 识别话题的首次报道,有助于 TDT 系统挖掘话题的种子事件并建立初始的话题模型,如话题检测系统可依赖种子事件建立话题的质心,以此作为识别和组织其它相关事件的参照系;而话题跟踪系统则可将种子事件作为跟踪后续话题发展的先验知识,以此限制跟踪过程的偏移现象.因此,NED 在 TDT 领域具有重要的理论研究意义.此外,新话题往往淹没于每日海量的信息流中,这一现象极大限制了人们及时掌握重要的新闻动态.尤其金融和股票市场的商情跟踪以及面向国家安全的国际军情和政治态势分析等领域,都需要一种精确高效捕捉重要新闻事件的机制.因此,NED 也具有重要的实用价值.

NED 研究体系主要包含如下三类分支:“报道-报道型”(简写为 SS)、“报道-聚类型”(简写为 SC)以及融合后的“报道-聚类-报道型”(简写为 SCS).基于 SS 的 NED 系统将待测报道与所有前期报道进行相关性匹配,如果相关度都小于特定阈值,则判定该报道为新话题.SS 的主要缺陷涉及两方面:一方面,一对一的报道匹配并不能恰当描述新旧话题的可区分性,相关于同一话题但侧重不同子话题的报道也可以具备较低相关度;另一方面,高昂的计算代价限制了 SS 的实用化.基于 SC 的 NED 系统则对所有前期报道进行聚类,将每个聚类作为不同旧话题的描述,并通过匹配待测报道与聚类的相关性进行新话题识别.SC 可以显著提高 NED 系统的效率,但相比于 SS 却获得较低的准确率[5-6].基于 SCS 的 NED 系统融合 SS 和 CS 各自的特点,将聚类作为话题的宏观描述,而聚类中的报道代表话题内的具体内容.检测过程中,待测报道与聚类的匹配用于快速查询最相近的旧话题,而待测报道与话题内各报道的匹配用于准确描述相关性.SCS 在不损失准确率的同时显著提高了 NED 系统的效率,但准确率基本等价于 SS[7],并未改进系统识别新话题的能力.

上述方法检测性能的不足主要体现在漏检率过高,尽管后续研究融入命名实体[8-11]和词链[12]等语言信息改进了现有 NED 系统的性能,但并未明显减少漏检现象.比如,将命名实体融入 SCS 的方法[7]在语料 TDT2 和 TDT3 的测试中取得了目前最佳成绩,但其性能最优时的漏检率为 0.4310,即遗漏了近似 43%的新话题.漏检现象源于 NED 系统将新话题的首次报道误判为旧话题的后续相关报道,其原因是现有方法将话题和报道作为非结构化的个体进行匹配,忽视了它们内部由不同子话题组成的多元化信息,往往待测报道仅仅与旧话题的冰山一角具有强相似性,则被 NED 系统归为一类.

针对上述缺陷,本文提出基于子话题分治匹配(Division Comparison of Subtopic, DCS)的 NED 策略.DCS 分别将报道和旧话题切分为多个子话题,并检验报道中与旧话题相关的子话题所占的比例:比例越低,则该报道作为新话题的概率越高.此外,DCS 还检验与待测报道各子话题强相关的旧话题的分布规律,越离散则待测报道越可能论述新话题.本文第 2 节介绍新事件检测领域的前人工作并进行问题分析;第 3 节简要论述 DCS 涉及的两项预处理,即文本描述和子话题切分方法;第 4 节重点论述 DCS 算法;第 5 节介绍实验环境和安排;第 6 节分析实验结果;第 7 节总结全文.

2 相关研究及问题分析

2.1 相关研究

NED 的早期研究来自于 Allan[2] 和 Papka[13] 等,其方法主要采用单遍聚类算法,首先提取特征建立待测报道的 Query 描述;然后与所有前期报道的 Query 进行匹配;如果匹配获得的相似度都低于阈值,则系统检测到新话题.Lam[14] 则对前期报道进行聚类,并将每个聚类看作一个旧话题,同时基于聚类建立话题的 Query 描述;在此基础上待测报道直接与所有聚类进行匹配,如果存在相似度高于阈值的聚类,则将待测报道嵌入其中并重新建立对应的 Query 描述,否则系统判定待测报道为新话题.NED 的后续研究以上述方法为框架尝试了两方面的改进,即充分利用新闻语料的特性和建立更好的文本表示形式.

利用新闻语料特性的主流方法是融入命名实体(Named entity, NE)识别技术.其中,Papka[13] 将语料全集中频繁出现的 NE 排除于文本描述之外,而赋予地点类 NE 四倍于其它特征的权重.Giridhar[11] 将报道描述成三种向量空间,分别是包含全部特征的向量、仅包含 NE 特征的向量和排除 NE 特征的向量,并对比了三种向量空间对 NED 系统的影响,实验验证 NE 对某些报道的新话题识别具有促进作用,但某些报道在没有 NE 参与的情况下效果更好.基于这一现象,Giridhar 将报道预先进行分类,并检验 NE 作用于哪些新闻类有助于 NED 系

统识别新话题. 在此基础上, 清华大学[7] 进一步细化 Giridhar 的工作, 详细区分了不同 NE 类别与各新闻类别的关联性. 比如, 加强人物类 NE 在“选举”类新闻报道中的权重, 对 NED 系统识别新话题更有促进作用.

在文本表示方面, Yang[15] 在划分前期报道类别的基础上, 只选择类中最优的相关报道对话题进行描述. Brants[16] 改进了增量式 TFIDF 权重计算方法, 并采用向量空间模型进行文本进行表示, 在此基础上利用 Hellinger 距离匹配文本相关性. Stocks[12] 结合词典信息(WordNet)和上下文信息挖掘报道中的词链, 并基于词链建立文本表示. 词链是一组语义上具有继承性的相关特征, 其理论基础来源于文本结构的凝聚假设[17], 即构成文本的词、短语和句子不是孤立存在, 而是趋向于围绕一个中心内涵进行组织与论述. 本文核心思想的提出也是源于这一假设, 即话题中的特征以种子事件为中心进行论述, 区别在于特征对种子事件的凝聚趋势并不直接, 而是首先凝聚于话题的不同子话题, 然后由子话题围绕种子事件进行论述.

2.2　问题分析

如前文所述, 话题不是单一结构体, 而是由多个子话题组成的有机整体, 其中每个子话题都是对某一相关事件的描述. 比如, 话题“金大中获得诺贝尔和平奖”除对种子事件的描述外, 还包含“金大中推动南北朝鲜和解”与“陈水扁致贺电并展望两岸关系”等子话题. 与此相似的是, 一篇新闻报道往往也包含多个子话题. 但现有 NED 系统将话题和报道描述为单一结构的“词包”, 并基于“词包”间的匹配衡量相关性. 而如 2.1 节所述, 词特征首先凝聚于子话题, 再以整个子话题为单位参与文本的组织. 因此, “词包”间的匹配不仅淡化子话题的特性, 并使子话题间形成互为噪声的局面, 从而误导新话题的识别. 下面基于两项例证解释这一负面影响的产生过程.

首先通过一项“拼图游戏”验证子话题特性在篇章理解中的作用. 假设该游戏包含 8 个图板, 其在 2×2 面积内的正确拼法如图 1 中的左图, 可将其类比于话题的正确描述; 对于不了解正确拼法的游戏人而言, 很容易错拼为右图的图案, 可类比于 NED 系统错误地理解话题的内涵. 造成错拼的原因在于 8 个图板拼接方式的多样性以及游戏人对正确拼法缺乏先验知识, 其类比于缺乏子话题划分的“词包”可组织为多种形式的话题结构, 以下将这一现象简称为“拼图效应”. 假设游戏预先提供如下规则, 即“灰色”图板必须以直角边进行拼接, “白色”图板直角边不能拼接, 则错拼的概率可显著降低, 其可类比于子话题的特性对正确理解话题的正面作用. 而“词包”因淡化子话题的特性, 削弱了 NED 系统理解话题的能力.

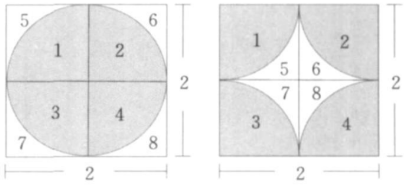

图 1　“拼图游戏”样例

现将上例映射到真实话题, 借以检验“拼图效应”造成子话题互为噪声现象的成因. 假设某一旧话题为“金大中获得诺贝尔和平奖”, 其子话题 i 和子话题 j 的特征空间如图 2 所示; 其中, 子话题 i 论述金大中获奖原因, 即“推动南北朝鲜和解”; 子话题 j 论述后续某一相关事件, 即“陈水扁致贺电并展望两岸关系”. 话题的真实含义中, 子话题内的特征具备更强的凝聚性, 而非结构化的“词包”破坏了这一凝聚性, 在所有特征之间建立了等价的联系. 在此基础上, “拼图效应”可将“词包”拼接为多种形式的话题结构, 如图 2 中列举的错拼方案 k, 其中包含一项“陈水扁推动两岸关系和解”的结构, 其内部的特征来源于不同子话题, 蕴涵的语义与原话题截然相反, 从而形成“词包”框架下子话题的特征互为噪声的现象. 由此产生的影响是, 该结构与题为“陈水扁侈谈推动两岸和解”的后续报道 s 具有强相关性, 而报道 s 是新话题的首次报道, 因此增大了 NED 系统漏检的可能性. 此外, 现有融入 NE 技术的方法在增强特征“陈水扁”权重的同时, 也增强了报道 s 与错拼结构的相关性, 反而助长了该例中的漏检趋势, 其有助于解释 Giridhar[11] 提出的 NE 技术在 NED 中性能不稳定的结论. 虽然 NED 系统在计算“词包”间相似度时, 并不存在结构化的拼接过程, 但实际匹配中却隐含了这一过程的影响, 一旦报道与话题的特征存在上例的匹配关系, 则 NED 系统间接包含图 2 中的流程.

基于子话题的分治匹配方法, 即 DCS, 仅在子话题间进行相关性匹配, 可有效屏蔽“拼图效应”引起的子话题互为噪声现象, 如图 2 中报道 s 与各子话题的相关度明显低于 s 与“词包”间的相关度, 利于降低 s 被误判为旧话题的概率. DCS 的另一优点是强化了子话题自身的特性, 如种子事件(“金大中获得诺贝尔和平奖”)与报道 s 的相关度明显低于子话题 j 与 s 的相关度, 可利用前者的这一特性削弱话题与 s 的关联性, 从而间接降低误判的可能性.

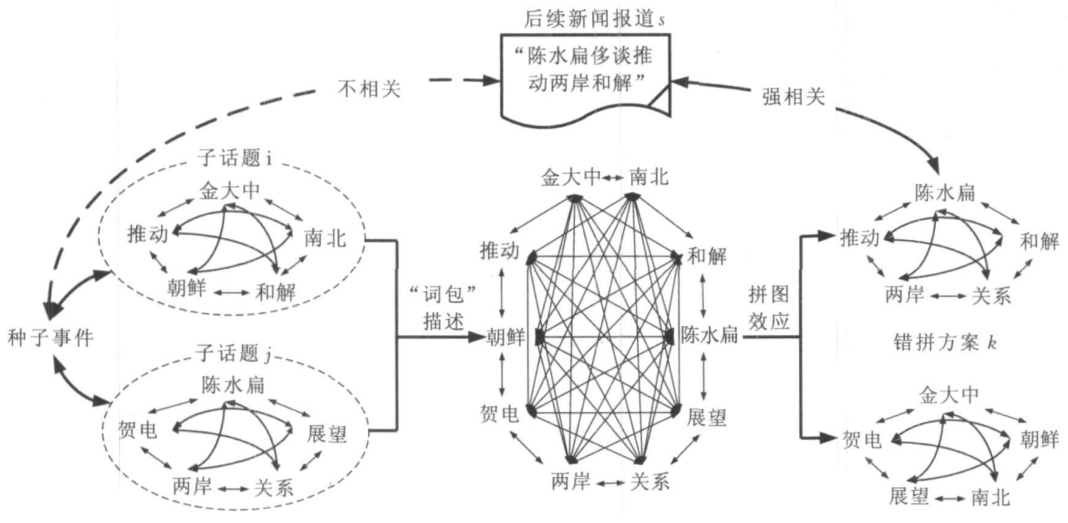

图 2 子话题互为噪声样例

3 预处理:文本描述及子话题划分

3.1 基于增量式 TFIDF 的文本描述

NED 针对实时报道流进行检测,报道之间具备严格的时序关系,对某一待测报道进行描述时,可利用的资源只局限于该报道之前的报道流. 基于这一特点,DCS 采用增量式 TFIDF 计算特征权重[5],并在此基础上建立文本的向量空间模型. 增量式 TFIDF 中,特征频率 tf 继承了传统计算方法,即特征在报道中出现的次数;而文档频率 df 则随时间动态更新,而不是基于全体语料训练获得的恒定值. 假设报道流按时序划分为多个时段,则特征 w 在某一时段 t 内的 df 计算如下式:

$$df_t(w)=df_{t-1}(w)+df_{D_t}(w) \tag{1}$$

其中,D_t 表示时段 t 内的报道集合;$df_{D_t}(w)$ 表示 D_t 内包含特征 w 的报道数量;$df_{t-1}(w)$ 表示时段 t 之前报道流中包含特征 w 的报道数量. 本文将 D_t 的规模设置为 50,即每个时段包含 50 篇报道[7]. 在此基础上,特征 w 的权重计算如下式:

$$weight(d,t,w)=\frac{\log(tf(d,w)+1)\times\log((N_t+1)/(df_t(w)+0.5))}{\sum_{w'\in d}\log(tf(d,w')+1)\times\log((N_t+1)/(df_t(w')+0.5))} \tag{2}$$

其中,$weight(d,t,w)$ 表示报道 d 内某一特征 w 在 t 时段的权重;N_t 表示时段 t 之前信息流包含的报道总数;$tf(d,w)$ 为特征 w 在报道 d 内的频率. DCS 采用向量空间模型对报道进行描述,其表示形式为 $d\rightarrow\{weight(d,t,w_1),weight(d,t,w_2),\ldots,weight(d,t,w_n)\}$,$n$ 为 d 内的特征总数;$weight(d,t,w)$ 如式(2).

3.2 基于 TextTiling 的子话题划分

DCS 涉及话题与报道的子话题划分,前者可通过各相关报道的子话题取并集获得,因此本节重点介绍报道的子话题划分方法. DCS 基于 TextTiling[18] 算法划分报道的子话题,其核心思想是通过句子间相关度的下降坡度识别子话题的边界. 首先,TextTiling 将报道划分为有序排列的句子集,每个句子采用向量空间模型进行描述,其中特征权重计算如 3.1 节. 然后在每两个句子之间设置子话题的候选边界,假设某一候选边界 b_i 位于句子 s_i 和 s_{i+1} 之间,则两句相关度的下降坡度计算如下:

$$Slope_{b_i}=\begin{cases}r(s_{i-1},s_i)+r(s_{i+1},s_{i+2}), & \text{if } r(s_{i-1},s_i)>r(s_i,s_{i+1})\cap r(s_{i+1},s_{i+2})>r(s_i,s_{i+1})\\ r(s_{i-1},s_i), & \text{else if } r(s_{i-1},s_i)>r(s_i,s_{i+1})\cap r(s_{i+1},s_{i+2})<r(s_i,s_{i+1})\\ r(s_{i+1},s_{i+2}), & \text{else if } r(s_{i-1},s_i)<r(s_i,s_{i+1})\cap r(s_{i+1},s_{i+2})>r(s_i,s_{i+1})\\ 0, & \text{else}\end{cases} \tag{3}$$

其中 $Slope_{bi}$ 为候选边界 b_i 处的下降坡度;r 表示句子间的相关度,其采用归一化的 Hellinger 距离进行计算[16]. 如果下降坡度高于阈值 σ,则 b_i 被确定为某一子话题的边界. 在此基础上,报道基于各边界划分为不同子话题. 阈值 σ 的计算如式(4),n 为报道内候选边界的总数;S 为所有候选边界下降坡度的均值.

$$\sigma=\frac{\sum_{i=1}^{n}(Slope_{b_i}-S)^2}{n} \tag{4}$$

4　子话题分治匹配算法

4.1　DCS 新话题识别

如第 2 节所述，子话题有助于新旧话题的区分，但第 2 节的分析过程建立在一个话题和一个报道之间，而 NED 系统需要在面向所有旧话题的情况下，对某一报道是否论述新话题进行判断. 对输入 NED 系统的待测报道，DCS 首先进行分词和去停用词处理，并在此基础上进行文本描述和子话题划分. 然后，针对该报道的每个子话题，DCS 查询一项最相关的旧话题，其查询流程如图 3 所示.

以图4中的待测报道 S_x 为例说明这一流程，DCS 从所有旧话题中查询出 T_i 的子话题 u'_{i1} 与 S_x 的子话题 u_{x1} 最相关，同时其相关度高于阈值 λ(λ=0.18，其训练过程如 6.1 小节)，则 DCS 将 T_i 作为 u_{x1} 最相关的旧话题；与此不同的是，DCS 虽然对 S_x 其它三个子话题也查询出最相关的匹配关系，但由于相关度低于阈值 λ，如 u'_{k1} 和 u_{x4} 的相关度为 0.14，因此 DCS 判定它们不相关于任何旧话题.

(1) 假设 u 为待测报道的某一子话题，DCS 遍历所有旧话题的子话题，从中查询最相关于 u 的子话题. 其中，子话题间的相关度计算采用 Hellinger 距离算法；

(2) 假设 u' 为步(1)中最相关于 u 的子话题，则 DCS 将两者的相关度与阈值 λ 进行比较，如果大于 λ，则 DCS 将 u' 所隶属的话题作为 u 最相关的旧话题；否则，u 不相关于任何旧话题.

图 3　DCS 查询子话题最相关的旧话题流程

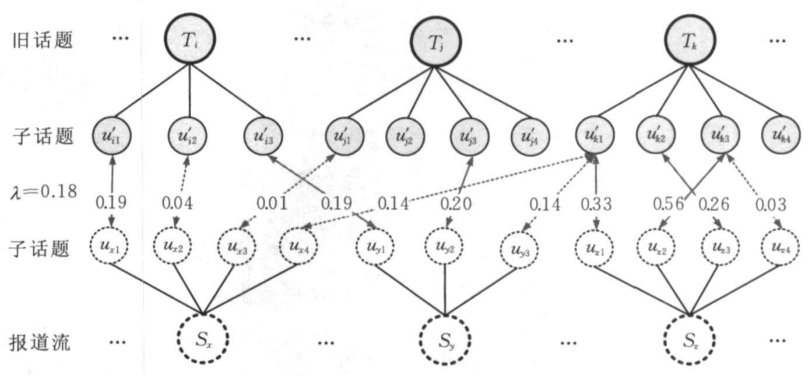

图 4　DCS 新话题识别样例

查询过程从两个角度检验了待测报道与旧话题的关系，分别是待测报道存在多少子话题相关于旧话题，简称为比例关系；以及待测报道的子话题相关于哪些旧话题，简称为分布关系，DCS 基于上述关系建立识别新话题的模型. 基于比例关系的建模过程中，DCS 计算相关于旧话题的子话题在待测报道中所占的比例，并将这一指标的倒数作为该报道论述新话题的概率，比例越小则概率越大，公式如下：

$$P_{\text{new}}(S)=\frac{\sum_{u\in S}u}{0.1+\sum_{u_r\in S}u_r} \tag{5}$$

其中，S 表示待测报道；u 表示 S 中的子话题；u_r 表示 S 中相关于旧话题的子话题；$P_{\text{new}}(S)$ 表示 S 论述新话题的概率. 该式中，分子统计 S 包含的子话题总数；分母统计 S 中 u_r 的数量，u_r 越少则 $P_{\text{new}}(S)$ 越高. 如图 4 中的报道 S_x，其子话题总数为 4，而其中只有一项子话题 u_{x1} 相关于旧话题，因此 S_x 论述新话题的概率较高. 该式为了平滑分母为 0 的情况，对其累加一项较小参数 0.1. 在式(5)的基础上，DCS 对 $P_{\text{new}}(S)$ 进行规一化并与阈值 θ 进行比较，如果高于 θ，则判定 S 论述新话题.

DCS 也尝试将分布关系融入识别新话题的建模过程，其核心思想是相关于某一旧话题的报道，其子话题也相关于该话题. 假设待测报道存在相关于旧话题的子话题，如果这些子话题趋向于凝聚，如图 4 中报道 S_z 的子话题相关于同一旧话题 T_k，则 DCS 削弱该报道论述新话题的概率；如果离散于不同旧话题，如图 4 中报道 S_y，则 DCS 强化该报道论述新话题的概率，其公式如下：

$$P'_{\text{new}}(S)=P_{\text{new}}(S)\cdot\frac{0.1+\sum_{T_r\in D}T_r}{0.1+\sum_{u_r\in S}u_r} \tag{6}$$

其中，$P'_{\text{new}}(S)$ 表示融入分布关系后报道 S 论述新话题的概率；D 为当前已知的所有旧话题；T_r 为 D 中与 S 的子话题相关的旧话题；u_r 表示 S 中相关于旧话题的子话题；$P_{\text{new}}(S)$ 如式(5). 式(6)中的分子部分统计 T_r 的数量，分母为 S 中 u_r 的数量，两者的比值描述了相关于 u_r 的旧话题平均数量，其指标介于 0 和 1 之间. 该平均值越高，则 u_r 的分布关系越离散，从而 $P_{\text{new}}(S)$ 相对被强化，反之则被削弱. 融入

分布关系的 DCS 进一步增强了新旧话题的可区分性,如图 4 中报道 S_x 的 $P'_{new}(S_x)$[①]等于 $P_{new}(S_x)$ · 1;而报道 S_z 的 $P'_{new}(S_z)$ 则为 $P_{new}(S_x)$ · 1/3,前者相对于后者论述新话题的概率被强化了,后者则相对被削弱,从而提高了 DCS 区分新旧话题的能力. 该式通过累加较小参数 0 1 平滑分子为 0 的情况.

4.2 DCS 旧话题更新

DCS 对输入 NED 系统的每篇报道进行检测,根据检测结果对旧话题进行更新,过程如图 5 所示.

假设 DCS 判定图 4 中的报道 S_x 论述新话题,而 S_y 和 S_z 相关于旧话题,则 DCS 的旧话题更新结果如图 6. 其中,DCS 基于 S_x 的子话题建立初始话题模型 T_{S_x},当后续报道输入 NED 系统时,T_{S_x} 作为旧话题参与检测过程;S_z 的子话题中除 u_{z4} 外都相关于旧话题 T_k(u_{z4} 与 T_k 的相关度低于 λ,判定为不相关),因此更新过程将 u_{z4} 嵌入 T_k 的话题模型;S_y 的子话题相关于两个旧话题,DCS 选择其中最相关的旧话题进行更新.

(1) 如果报道 s 论述新话题,则 DCS 将 s 作为种子,建立话题的初始模型 T_s,T_s 由 s 的所有子话题进行描述. 从下一篇进入 NED 系统的报道开始,T_s 将作为旧话题参与后续新话题的识别;

(2) 如果 s 相关于某一旧话题 T_r,则 DCS 检验 s 中是否存在不相关于 T_r 的子话题,如果存在,则将该子话题嵌入 T_r.

图 5 DCS 旧话题更新流程

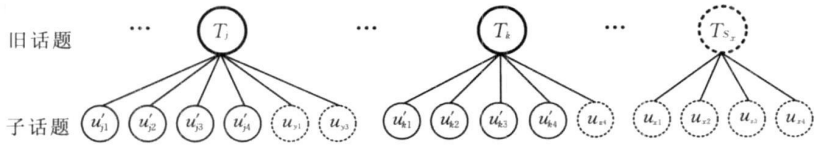

图 6 DCS 旧话题更新样例

针对子话题相关于多个旧话题的情况,4 1 节尝试利用分布关系强化其描述新话题的概率,但如果相关于旧话题的子话题在报道中占有很高比例,则基于式(6)计算出的概率仍然偏低,即报道相关于旧话题. 此时,报道的相关性离散于不同旧话题,如图 4 中的 S_y,因此更新过程需要选择更新对象,其首先计算该报道与每个相关旧话题的相关度,公式如下:

$$Sim(S, T_r)=\sum_{u_r \in S} Sim(u_r, T_r) \tag{7}$$

其中,Sim 表示相关度;其它各符号的含义与式(6)相同. 该式选取 S 中相关于 T_r 的所有子话题,统计它们与 T_r 相关度的和,并将该指标作为 S 与 T_r 的相关度. 在此基础上,DCS 选择与 S 相关度最高的旧话题进行更新. 因此,图 4 中的 S_y 对应的更新对象为 T_j,更新过程将 S_y 中不相关于 T_j 的子话题 u_{y1} 与 u_{y3} 嵌入 T_j 的话题模型,如图 6.

5 实验设计

5.1 语 料

实验采用语言数据协会(LDC)提供的 TDT4 和 TDT5 中文语料进行评测. TDT 语料由路透社和 CNN 等媒体机构播报的新闻报道组成,包含中文、英文和阿拉伯文三种语言,并涉及文本、音频广播及其翻录三种表述方式. 该实验基于文本形式的中文语料进行训练与测试,其中 TDT4 提供了针对 100 个话题的 1303 篇中文报道,TDT5 提供了针对 59 个话题的 999 篇中文报道.

5.2 评 测

实验基于美国国家标准与技术研究院(NIST)针对 TDT 发布的评测指南,采用检测错误代价 C_{Det} 分别从漏检和误检两个角度进行评测,公式如下:

$$C_{Det}=C_{Miss}P_{Miss}P_{target}+C_{FA}P_{FA}P_{non-target} \tag{8}$$

其中,P_{Miss} 和 P_{FA} 分别表示系统的漏检率和误检率,漏检即为系统未识别出新话题,误检则是系统将旧话题的后续相关报道误判为新话题;C_{Miss} 和 C_{FA} 分别代表漏检和误检的代价系数($C_{Miss}=1$, $C_{FA}=0.1$);P_{target} 和 $P_{non-target}$ 是先验目标概率($P_{target}=0.02$, $P_{non-target}=1-P_{target}$). 检测错误代价 C_{Det} 的规范化形式 $Norm(C_{Det})$ 如式(9). 此外,NIST 面向 TDT 研究提供了可视化的评测工具,即检测错误权衡图(Detection Error Tradeoff, DET). DET 利用二维坐标系的纵轴表示系统漏检率;横轴表示误检率,并根据漏检与误检随阈值 θ 的变化趋势绘制系统的性能曲线. 由于系统漏检与误检的概率越低,其性能越好,因此 DET 曲线越靠近坐标系的左下角代表系统性能更优. DET 曲线上的最小 $Norm(C_{Det})$ 指标代表检测系统的最佳性能,简写为 Min $Norm(C_{Det})$.

$$Norm(C_{Det})=\frac{C_{Det}}{\min(C_{Miss}\cdot P_{target}, C_{FA}\cdot P_{non-target})} \tag{9}$$

5.3 实验安排

实验分为三个组成部分,首先训练子话题相关

① 待测报道 S 中相关于旧话题的子话题数为 0 或 1 时,其不存在分布关系,式(6)的计算结果维持了原有 $P_{new}(S)$,相对于其它因分布关系趋向于凝聚而被削弱的报道,S 论述新话题的概率被强化.

性匹配的阈值 λ；其次，分别建立如下系统并比较性能：

(1) SCS＋NE：系统采用融合命名实体的“报道-聚类-报道型”算法，SCS 如第 1 节，NE 如 2.1 小节[7]；

(2) DCS(P)：系统采用基于比例(Proportion)关系的 DCS 算法，其新话题识别如式(5)；

(3) DCS(P＋D)：系统采用比例关系和分布(Distribution)关系相融合的 DCS 算法，如式(6).

其中，(2)和(3)的旧话题更新算法相同，如 4.2 小节；SCS＋NE 在现有 NED 系统中性能最佳[7]，实验将其作为对比测试的 Baseline，由于其公布的性能指标来源于对英文语料的测试，而本实验针对中文语料进行 NED 检测，因此实验重新实现 SCS＋NE 算法，并与 DCS 统一于 TDT4 和 TDT5 中文语料进行测试. 最后，实验分析并对比 DCS 与 SS、SC 以及 SCS 的检测效率.

6　实验结果与分析

6.1　阈值 λ 估计

实验基于 TDT4 中文语料，通过观察 $Norm(C_{Det})$ 随阈值 λ 的变化趋势，实现最佳 λ 值的估计，如图 7. 其中，纵轴表示 $Norm(C_{Det})$；横轴表示新话题识别中的阈值 θ；二维坐标系内的曲线表示 λ 取某一指标时，$Norm(C_{Det})$ 随 θ 的变化趋势；不同曲线间的差异描述 $Norm(C_{Det})$ 随 λ 的变化趋势. 图中曲线越靠下，则系统的检测性能越好，即花费的检测错误代价 $Norm(C_{Det})$ 越小. 基于这一训练过程，实验将阈值 λ 设置为 0.18. 图 7 显示不同 λ 对应的检测性能迥异，其侧面反映了子话题间的相关性是区分新旧话题的重要因素，从而验证了子话题分治策略的可行性.

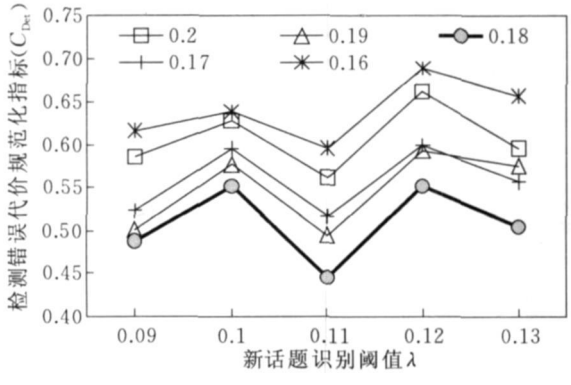

图　7

6.2　检测性能对比与分析

实验分别建立系统 DCS(P)、DCS(P＋D)以及 SCS＋NE，如 5.3 节，并对比它们的检测性能. 图 8 是上述系统基于 TDT4 中文语料获得的 DET 对比图，其横轴表示误检率；纵轴表示漏检率；二维坐标系内的曲线代表系统识别新话题的整体性能；曲线上的点表示某一阈值 θ 对应的一组误检和漏检指标；图中三个几何形状的标识分别对应不同系统的最小检测错误代价，即 Min $Norm(C_{Det})$. 如图所示，与系统 SCS＋NE 相比，系统 DCS(P)与 DCS(P＋D)的 DET 曲线更趋近于二维坐标系的左下角，说明其误检与漏检的概率更低；与此同理，DCS(P＋D)的检测性能略优于 DCS(P). 此外，DCS(P)与 DCS(P＋D)的 Min $Norm(C_{Det})$ 分别为 0.4464 和 0.4061，相比于 SCS＋NE 的 0.5050 分别改进了 5.9%和 9.9%；同理 DCS(P＋D)相比于 DCS(P)改进了 4%. 上述改进主要来源于漏检率的降低，DCS(P)与 DCS(P＋D)在 Min $Norm(C_{Det})$ 处的漏检率分别为 0.2196 和 0.1859，远低于 SCS＋NE 的 0.4，此时 DCS(P)与 DCS(P＋D)的误检率仅比 SCS＋NE 增加了 0.025 和 0.024.

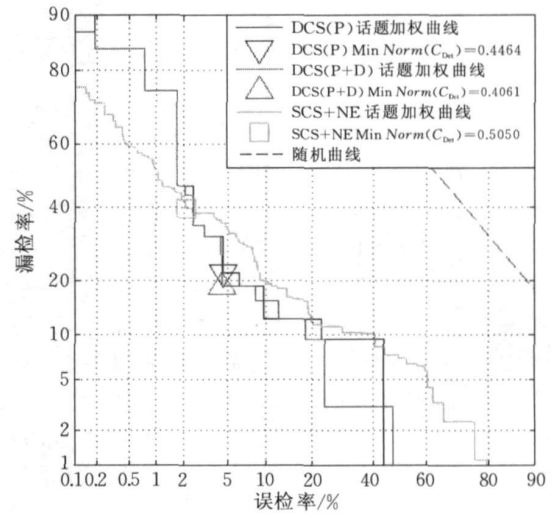

图 8　基于 TDT4 获得的 DET 对比图

该实验结果验证了子话题分治策略对新话题识别的有效性，其削弱了不同子话题间的互为噪声现象以及子话题独立语义被泛化现象对 NED 系统的负面影响. 该结果同时也验证融入分布关系的 DCS(P＋D)显著改进了 DCS(P)的检测性能，原因在于分布关系细化了 DCS 基于比例关系获得的相关性粒度，如图 8 中 DCS(P＋D)的 DET 曲线比 DCS(P)具有更多性能变化的折点，此外分布关系也强化了新旧话题的可区分性，从而使阈值 θ 可更准确地定位划分新旧话题的边界. 尽管如此，DCS 的相关性粒度仍然很稀疏，如图 8 中 DCS(P)与

DCS(P+D)的性能变化折点远少于SCS+NE,原因是报道包含的子话题数量很少,TDT4 语料中报道的子话题数量最大值为 9,基于式(5)获得的相关度最多为 45 种情况,而大部分报道的相关度仅局限于其中几种.这一现象的优点是 NED 系统可以相对便捷地训练和设置阈值,缺陷在于系统性能对阈值的敏感性增强了,如图 8 中 DCS(P)与 DCS(P+D)的 DET 曲线在两端的性能落差很大.这一缺陷更直观地体现在基于 TDT5 的测试结果,如表 1.其中 $Norm(C_{Det})$ 是各系统采用 TDT4 最佳阈值 θ 在 TDT5 中获得的检测性能;Min $Norm(C_{Det})$ 是各系统在 TDT5 中可达到的最佳性能.表 1 显示 SCS+NE 在不同语料中采用同一阈值对系统性能影响不大,而 DCS(P)与 DCS(P+D)则因系统性能对阈值过于敏感,使 TDT4 训练出的最佳阈值在 TDT5 中并不适用,其获得的检测性能与系统在 TDT5 中可达到的最佳性能落差很大.

表 1 基于 **TDT5** 的性能对比

	$Norm(C_{Det})$	$Min\ Norm(C_{Det})$
SCS+NE	0.5168	0.4986
DCS(P)	0.4954	0.4396
DCS(P+D)	0.4522	0.4004

6.3 效率分析

实验选择三种主要 NED 方法,即 SS、SC 和 SCS,与 DCS 的效率进行比较.效率测试将匹配次数和每次匹配涉及的特征数量作为衡量标准.假设现有旧话题的数量为 N,已检测过的报道数为 n(分别相关于 N 个旧话题),每个报道包含的平均子话题数为 ε,每个子话题包含的平均特征数为 μ,则上述方法的效率分析如下:SS 将待测报道与所有已检测的报道进行匹配,其匹配次数为 n,每次匹配涉及的特征数为 $\varepsilon\mu$,因此 SS 的时间复杂度可标记为 $O(SS)=n\varepsilon\mu$;SC 将待测报道与每个旧话题进行匹配,其匹配次数为 N,由于 SC 将相关报道的质心作为话题的特征向量,对其中不同报道间重复的子话题,质心计算主要进行权重更新,而新特征的嵌入主要来自新的子话题,因此话题质心由不同子话题的特征组成,从而 SC 的复杂度可标记为 $O(SC)\approx N(\varepsilon+\gamma)\mu$,其中 ε 表示旧话题首次报道包含的子话题数,γ 表示后续相关报道中新的子话题数,又因为 $\gamma<n/N$,即话题内包含的子话题少于相关于该话题的报道,且 $N\ll n$,因此 $O(SC)\ll n(\varepsilon+1)\mu\approx O(SS)$;SCS 则首先广度遍历所有旧话题,选择最相关于待测报道的话题,再深度遍历该话题下的所有报道,查询其中最相关的报道参与新话题识别,因此其复杂度略高于 SC;而 DCS 将待测报道与每个旧话题的子话题进行匹配,其复杂度为 $O(DCS)=N(\varepsilon+\gamma)\mu$,近似于 SC.基于上述分析,现有方法与 DCS 的效率评价为 $O(SC)\approx O(DCS)<O(SCS)\ll O(SS)$.实验中,DCS 的平均匹配时间约为 SC 的 1.04 倍,SCS 的 0.44 倍.

7 结论与展望

本文首先分析了传统单一文本结构在新话题识别过程中的缺陷以及基于子话题将报道和话题描述为多元结构的优点.在此基础上,本文提出基于子话题分治匹配(DCS)的新事件检测算法,并在 TDT4 和 TDT5 中文语料上进行实验,其结果验证该算法显著改进了现有方法漏检率过高的不足.

DCS 仅是利用子话题解决新事件检测问题的初步尝试,其仍存在如下缺陷:

(1)无论报道或话题都至少存在某一子话题论述种子事件,这类子话题在区分新旧话题中的价值更高,而 DCS 在子话题间形成了等价关系;

(2)某些子话题可隶属于不同话题,其不利于新话题的识别.因此,未来工作将尝试基于子话题建立层次化的话题模型,根据子话题间不同层次的关联程度分配它们在新话题识别中的贡献度;在此基础上,未来工作还将基于时序屏蔽过早的旧话题参与后续新话题的识别,从而进一步提高系统检测效率.

参 考 文 献

[1] Luo Wei-Hua, Liu Qun, Cheng Xue-Qi. Development and Analysis of Technology of Topic Detection and Tracking//Proceedings of the JSCL-2003. Beijing: Tsinghua University Press, 2003: 560-566(in Chinese)
(骆卫华,刘群,程学旗.话题检测与跟踪技术的发展与研究//全国计算语言学联合学术会议(JSCL-2003)论文集.北京:清华大学出版社,2003:560-566)

[2] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking//Proceedings of the SIGIR'98: 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1998: 37-45

[3] Allan J. Topic Detection and Tracking: Event-based Information Organization. USA: Kluwer Academic Publishers, 2002: 1-16

[4] Yang Y, Carbonell J, Brown R et al. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 1999, 14(4): 32-43

[5] Yang Y, Pierce T, Carbonell J. A study on retrospective and on-line event detection//Proceedings of the SIGIR-98. Melbourne, Australia, 1998: 28-36

[6] Allan J, Lavrenko V, Malin D et al. Detections, bounds, and timelines: Umass and TDT-3//Proceedings of the Topic Detection and Tracking Workshop (TDT-3). Vienna, VA, 2000: 167-174

[7] Zhang K, Li J, Wu G. New event detection based on indexing-tree and named entity//Proceedings of the SIGIR-2007. Ámsterdam, Netherlands, 2007: 215-222

[8] Yang Y, Zhang J, Carbonell J et al. Topic conditioned novelty detection//Proceedings of the 8th ACM SIGKDD International Conference. Edmonton, Canada: ACM Press, 2002: 688-693

[9] Juha M, Helena A M, Marko S. Applying semantic classes in event detection and tracking//Proceedings of the International Conference on Natural Language Processing (ICON 2002). Turku, Finland, 2002: 175-183

[10] Juha M, Helena A M, Marko S. Simple semantics in topic detection and tracking. Information Retrieval, 2004, 7(3-4): 347-368

[11] Giridhar K, Allan J. Text classification and named entities for new event detection//Proceedings of the 27th Annual International ACM SIGIR Conference. New York, USA: ACM Press, 2004: 297-304

[12] Nicola S, Joe C. Combining semantic and syntactic document classifiers to improve first story detection//Proceedings of the 24th Annual International ACM SIGIR Conference. New York, USA: ACM Press, 2001: 424-425

[13] Papka R, Allan J. On-line new event detection using single pass clustering TITLE2: Technical Report UM-CS-1998-021, 1998

[14] Lam W, Meng H, Wong K et al. Using contextual analysis for news event detection. International Journal on Intelligent Systems, 2001, 16(4): 525-546

[15] Yang Y, Pierce T, Carbonell J. A study on retrospective and on-line event detection//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia: ACM Press, 1998: 28-36

[16] Thorsten B, Francine C, Ayman F. A system for new event detection//Proceedings of the 26th Annual International ACM SIGIR Conference. Toronto, Canada: ACM Press, 2003: 330-337

[17] Croft W B, Townsend S C, Lavrenko V. Relevance feedback and personalization: A language modeling perspective//Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries. Dublin, Ireland, 2001: 49-54

[18] Marti A H. Multi-paragraph segmentation of expository text//Proceedings of the 32nd Annual Meeting of the ACL. Las Cruces, New Mexico, 1994: 9-16

HONG Yu, born in 1978, Ph. D. candidate. His research interests include topic detection and tracking, information filtering, personal information retrieval.

ZHANG Yu, born in 1972, Ph. D., associate professor. His research interests include information filtering, automatic question and answer, natural language processing.

FAN Ji-Li, born in 1986, M. S. candidate. His research interests focus on topic detection and tracking.

LIU Ting, born in 1972, Ph. D., professor, Ph. D. supervisor. His research interests include natural language processing, information retrieval.

LI Sheng, born in 1943, Ph. D., professor, Ph. D. supervisor. His research interests include natural language processing, information retrieval, machine translation.

Background

Topic Detection and Tracking, named TDT for short, is a challenging direction of research on natural language processing which aims at developing technologies for event-based information organization, such as detecting stories reported on new topics and tracking stories on known topics. Linguistic Data Consortium, named LDC for short, provided multiple sources of information for training and test of TDT. These sources, including both text and speech, are namely newswires, radio and television news broadcast programs and WWW sources. The source languages are English, Mandarin and Arabic. The information streams are modeled as a sequence of stories, which provide information on many topics. National Institute of Standards and Technology, named NIST for short, provided guides, tools and dry runs for evaluation of TDT.

In the initial TDT research, conducted during 1996 and 1997, the notion of a topic was limited to be an "event", meaning something that happens at some specific time and place. In the second TDT project, TDT2, the definition of a topic was a seminal event or activity, along with all directly related events and activities. This definition was retained for the third project, TDT3, TDT4 and TDT5. A story will be considered to be "on topic" whenever it discusses events and activities that are directly connected to that topic's seminal event. TDT includes many tasks, such as story segmentation, link detection, topic detection, new event detection, topic tracking, etc. New event detection, named NED for short, is one of the most important tasks in TDT, which aims at real-time monitoring a chronologically ordered stream of stories and identifying the first story that discussed an event.